近 5 年信息检索的研究热点与发展趋势综述*

——基于相关会议论文的分析

杨超凡 邓仲华 彭 鑫 刘 斌
(武汉大学信息管理学院 武汉 430072)

摘要:【目的】统计近 5 年相关会议集收录的论文, 分析信息检索的研究热点与发展趋势。【文献范围】检索 2012 年–2016 年 ACL、ACMMM、ICML、KDD、SIGIR 等 5 个信息检索领域的相关会议集收录的论文。【方法】使用爬虫软件获取 5 个相关会议收录的论文的摘要和关键词, 并利用分词工具对其处理, 进行统计分析和文献研究。【结果】发现目前信息检索中移动搜索是主流; 检索模型不断优化; 注重过滤和推荐; 与人工智能关系密切, 用户隐私以及医疗健康也是信息检索重点关注的内容。【局限】仅采集论文的摘要和关键词数据, 未进行全文内容以及引文的分析。【结论】反映目前信息检索的大致发展状况, 为其他学者开展新的研究提供借鉴和参考。

关键词: 信息检索 会议论文 研究热点 发展趋势
分类号: G250

1 引 言

近年来, 国内外学者对信息检索领域的研究成果相当丰富。司莉等以 WoS、ACM、Emerald、Elsevier、ProQuest、Springer 等数据库收录的文献为基础, 对近 10 年来多语言信息组织与检索的研究进行述评[1]; 吴丹等以国外期刊发表的协同信息检索行为研究文献为对象, 采用综合归纳方法, 分析协同信息检索行为研究的进展[2]; 杨海锋跟踪国内外重要研究成果, 对用户行为在信息检索中的研究现状进行了概况和评述, 并总结了信息检索面临的诸多挑战[3]; 窦永香等对 ACM SIGIR 年会进行主题分析, 总结信息检索的主要研究内容、研究热点及最新研究动向[4]; 陈少涌等采用文献计量和社会网络分析等方法对近 10 年来 ACM SIGIR 年会的主题及论文进行统计分析, 揭示了信息检索领域在过去 10 年的文献主题分布、作者分布情况[5]。

国外学者的研究则更倾向于对信息检索的具体应用的综述。Casey 等探讨了以内容为基础的音乐信息检索目前的发展方向以及未来面临的诸多挑战[6]; Kishida 对当前最先进的跨语言信息检索技术和方法进行综述[7]; Enser 援引大量文献综述了数字化时代的图像检索、视频检索、语义图像检索等可视化信息检索的发展历程[8]; Smeaton 等通过对 SIGIR 过去 25 届年会的主题和合著情况作分析, 概要地展示了 SIGIR 25 年中不同主题的分布及作者发文情况[9]; Hiemstra 等对过去 30 年中 SIGIR 会议文献作分析并揭示了研究主题、作者分布及合著情况[10]。

综合来看, 国内外学者大多以相关期刊论文为对象分析信息检索的研究热点和发展趋势, 少部分研究采用会议论文进行分析, 且在这些研究中作者均选择了 ACM SIGIR 这一会议的论文为研究对象, 鉴于目

通讯作者: 杨超凡, ORCID: 0000-0001-6327-9924, E-mail: yangchaofan@whu.edu.cn。

*本文系国家自然科学基金项目“大数据环境下面向科学研究第四范式的信息资源云研究”(项目编号: 71373191)和国家自然科学基金项目“云计算环境下图书馆的信息服务等级协议研究”(项目编号: 71173163)的研究成果之一。

前信息检索的发展方向十分丰富，综合多个相关国际会议论文进行对比分析就显得更有理论价值和实践意义。

2 数据来源与研究方法

本文选取国际计算机语言学协会年会(ACL)、国际计算机学会多媒体会议(ACMMM)、国际机器学习大会(ICML)、国际计算机学会知识发现与数据挖掘年会(KDD)以及信息检索特别兴趣小组会议(SIGIR)等 5 个信息检索相关的国际会议 2012 年–2016 年收录的论文为对象，通过网络爬虫工具获取其摘要及关键词，并利用 Stanford CoreNLP 进行分词处理[11]，进行研究趋势的统计和归类分析。

本文主要运用以下研究方法:

(1) 网络调查法，通过 5 个国际会议的官方网站查找和获取有关的论文数据，使得本文比较具有真实性。

(2) 文献研究法，借助 ACM DL、WoS、中国知网等数据库检索有关文献并发现其可参考之处，使得本文更加全面和具有说服力。

(3) 统计分析法，使用网络爬虫工具获取了 2012 年–2016 年 5 个国际会议出版论文的数量及其摘要和关键词，对热点词的词频进行统计，从而分析信息检索领域的研究趋势，因此本文具有较强的严谨性。

3 总体统计与分析

各年度会议论文数量的变化趋势如图 1 所示，总体来看，5 个会议在 2012 年–2016 年收录的论文数都呈上升的趋势，信息检索的研究和应用不断丰富，推动着相关会议接收论文数的持续增长。

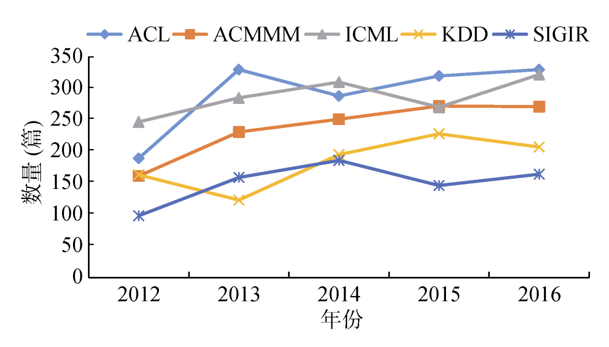


图 1 各年度会议论文数量变化趋势

对 5 个会议 2012 年–2016 年收录论文的摘要做分词处理，并与关键词一起在 Excel 中进行筛选和排序，得到排名前 20 的热点词，利用 WordArt.com 生成词云图，如图 2 所示，可以发现这些热点词集中在以下几个方面:

(1) 信息检索类型: 图像检索(image retrieval)、视频搜索(video search)以及音乐搜索(music search)等。

(2) 信息检索技术: 事件检测(event detection)、协同过滤(collaborative filtering)、特征选择(feature selection)和矩阵分解(matrix factorization)等。

(3) 信息检索研究情境: 社会网络(social networks)以及社交媒体(social media)。

(4) 信息检索关联技术: 自然语言处理(natural language processing)、机器学习(machine learning)、深度学习(deep learning)以及神经网络(neural networks)等。



图 2 各会议 2012 年–2016 年热点词词云图

4 研究热点分析

4.1 热点词统计排名分析

对各年度 5 个会议的热点词分别进行了统计和排序，得到各年度会议论文热点词频次排名如表 1 所示。由此得出以下发现:

(1) 人工智能相关的热点词词频均处于上升趋势，其中神经网络和深度学习在 5 年间迅速成为信息检索的重点研究内容。神经网络是人工神经网络(Artificial Neural Network)的简称，人工神经网络是从信息处理角度对人脑神经元网络进行抽象，建立某种简单模型，按不同的连接方式组成不同的网络[12]，其中递归神经网络(Recurrent Neural Network, RNN)和卷积神经网络(Convolutional Neural Network, CNN)是学者们关注的焦点。RNN 是一种反馈神经网络，是一个非线性动力系统，可用来实现联想记忆和求解优化等问题[13];

CNN 是一种前馈神经网络, 它的人工神经元可以响应一部分覆盖范围内的周围单元[14], 其主要用于处理大型图像。深度学习是机器学习中一种基于数据进行表征学习的方法, 其主要优势是利用非监督式或半监督式的特征学习和分层特征提取高效算法来替代手工获取特征[15]。

(2) 相对而言, 学者们对一些信息检索技术的关注度小幅下降, 如特征选择和矩阵分解等热点词的出现频率整体上逐年减少。特征选择是从原始特征中选择出一些最有效特征以降低数据集维度的过程, 是提高学习算法性能的一个重要手段, 也是模式识别中关键的数据预处理步骤[16]; 矩阵分解的思想就是信息检索中用户和物品都会有一些特性, 矩阵分解可以从评分矩阵中分解出用户——特性矩阵, 特性——物品矩阵, 从而得到了用户的偏好和每件物品的特性以及确定矩阵的维度[17]。由此看出目前信息检索研究已经不只局限于自身相关的技术, 而是转向融合其他学科技术的综合研究。

表 1 各年度会议论文热点词频次排名表

序号	热点词	总词频	各年度热点词频次				
			2012	2013	2014	2015	2016
1	神经网络	394	9	19	81	110	175
2	机器学习	386	45	72	72	87	110
3	社交网络	379	109	81	56	83	50
4	社交媒体	358	66	76	51	71	66
5	搜索引擎	315	73	70	69	79	24
6	信息检索	196	35	41	65	55	39
7	数据挖掘	148	43	31	31	28	15
8	图像检索	128	41	28	22	24	11
9	自然语言处理	126	6	10	7	52	51
10	主题模型	112	22	48	13	13	26
11	监督式学习	109	30	20	22	16	21
12	网页搜索	101	11	30	29	22	10
13	推荐系统	100	20	25	23	20	12
14	深度学习	88	5	5	11	18	49
15	视频搜索	88	27	20	17	14	10
16	事件检测	87	14	18	6	28	21
17	音乐搜索	86	11	14	15	21	24
18	协同过滤	82	14	22	19	16	11
19	特征选择	75	27	14	14	12	8
20	矩阵分解	75	21	22	16	10	6
21	主动学习	73	26	21	13	7	6
22	情感分析	68	16	8	10	14	20
23	语言模型	67	17	19	10	12	9
24	分词技术	65	6	14	20	13	12
25	增强学习	63	18	7	9	6	23

4.2 各会议主题分析

本文对各年度会议论文的主题进行整理和分类, 归纳出各年度会议论文主题词表, 如表 2 所示。从中可以看出:

(1) ACL 会议中机器翻译、信息抽取、问答系统以及自然语言处理等研究主题较为稳定, 但其中某些同一主题下的论文的研究内容也随时间的推进而发生演变。例如对信息抽取的关注点从语义关系抽取转向命名实体识别和大规模信息抽取; 在机器翻译方面, 从基于短语的统计机器翻译研究转向端对端的神经机

表 2　各年度会议论文主题词表

年度 会议　主题	2012	2013	2014	2015	2016
ACL	机器翻译; 数据挖掘; 信息抽取; 问答系统; 文本分类; 自然语言处理应用	观点挖掘; 机器翻译; 自然语言处理应用; 问答系统; 机器学习; 文本分类; 信息抽取;	机器翻译; 自然语言处理; 分词技术与词性标注; 情感分析; 机器学习; 问答系统	神经网络; 机器学习; 信息抽取; 机器翻译; 问答系统; 自然语言处理; 主题模型	问答系统; 信息抽取; 神经网络; 机器翻译; 深度学习; 语义分析; 情感分析; 文本分类;
ACMMM	多媒体推荐; 持续性情感分析; 基于内容的图像检索; 大规模搜索; 人脸识别; 社交媒体	行为与事件识别; 多峰分析; 社会动力学; 相似性搜索; 情境感知; 音乐与戏剧分析	行为与事件识别; 深度学习; 人机交互; 多媒体分析与挖掘; 隐私与健康; 多媒体推荐; 移动搜索	多媒体标引与搜索; 行为与事件识别; 多媒体质量感知; 人机交互; 虚拟现实与增强现实; 移动设备	人脸与情感识别; 视频搜索; 深度学习; 虚拟现实与增强现实; 隐私与健康; 人机交互
ICML	聚类分析; 增强学习; 神经网络与深度学习; 优化算法; 隐私与保密; 监督式学习; 概率模型	增强学习; 深度学习; 社交网络; 主题模型; 支持向量机与决策树; 聚类分析; 优化算法; 矩阵分解	深度学习; 增强学习; 结构化预测; 聚类分析; 特征选择; 神经网络; 矩阵分解; 主题模型	深度学习; 概率模型; 增强学习; 结构化预测; 时间序列分析; 特征选择; 隐私研究; 聚类分析	神经网络与深度学习; 增强学习; 矩阵分解; 大数据; 监督式学习; 隐私研究; 图解模型; 聚类分析
KDD	网页级别与社交媒体; 模式挖掘; 概率模型; 监督式学习; 网站应用; 个性化推荐	文档与主题模型; 社交媒体; 大数据框架; 图像挖掘; 医疗与生活; 深度学习; 推荐系统	医疗与安全; 监督式学习; 社交媒体; 特征选择; 文本挖掘; 隐私与保密; 主题模型; 移动设备	大数据; 主题模型; 隐私与保密; 移动设备; 知识发现; 医疗健康; 模式挖掘; 推荐系统; 电子商务	图像与社交网络; 深度学习; 聚类分析; 推荐系统; 用户行为模型; 优化算法; 电子商务
SIGIR	多媒体; 检索评价; 推荐系统; 搜索日志分析; 社交媒体; 个性化与用户模型; 搜索效率; 文本分类	社交媒体; 推荐系统; 主题模型; 多媒体检索; 用户行为; 文本分类; 电子商务; 相似性搜索; 移动搜索	社交媒体; 移动搜索; 标引与搜索效率; 用户与模型; 情感分析; 引用推荐; 搜索满意度; 搜索风险评估; 哈希算法	多媒体搜索; 搜索体验; 社交媒体; 用户模型; 分类与排名; 深度学习; 任务与设备; 电子商务; 移动搜索	检索模型; 音乐与数学; 隐私、广告与产品; 行为模型与应用; 移动设备; 实体与知识图谱; 问答系统; 多媒体搜索

器翻译研究; 而对于问答系统的研究则由基于增强词汇的语义模型向分层并行的知识理解模型转变。总的来说, ACL 会议主要关注信息检索中的信息的获取与处理和语义的分析与理解等内容。

(2) ACMMM 会议的主题则较为丰富, 其中涉及到较多与信息检索相关的新技术。人机交互是指人与计算机之间使用某种对话语言, 以一定的交互方式, 为完成确定任务的人与计算机之间的信息交换过程[18], 有关人机交互方面的研究从行为与事件识别转向人脸识别以及情感识别。近两年, 虚拟现实与增强现实也成为 ACMMM 会议的热门研究主题, 虚拟现实技术是一种可以创建和体验虚拟世界的计算机仿真系统[19]; 增强现实技术是一种实时地计算摄影机影像的位置及角度并加上相应图像、视频、3D 模型的技术[20]。综合其他主题发现, ACMMM 会议主要呈现了新兴技术与信息检索的交互应用以及多媒体信息的搜索、推荐和评价。

(3) ICML 会议则较为专注于研究聚类分析以及深度学习、增强学习和监督式学习等机器学习的子类方法。聚类分析是将物理或抽象对象的集合分组为由类似的对象组成的多个类的分析过程, 其研究内容包括预测药物不良反应事件[21]以及子空间分割的多任务学习[22]等。综合来看, ICML 侧重于探索机器学习与信息检索的融合以及对检索模型的研究。

(4) KDD 会议的主题倾向于各类数据的挖掘与知识发现, 挖掘形式从文本挖掘、图像挖掘到模式挖掘, 研究的情境从 2012 年–2014 年的社交媒体到近两年的电子商务, 研究内容覆盖了网站应用、医疗健康、隐私与保密以及用户行为等。此外, 近 5 年 KDD 会议对推荐系统的研究也较为热门, 推荐系统是根据用户的兴趣特点和购买行为的过程, 向用户推荐用户感兴趣的信息和商品, 其推荐方法包括基于内容推荐、基于关联规则推荐、基于知识推荐和组合推荐[23]。从以上主题中可以看出, 信息检索与数据挖掘息息相关, 两者的综合应用前景广阔。

(5) SIGIR 会议的研究主题大多集中在信息检索技术以及应用，且研究内容由搜索日志分析、文本分类以及检索模型转向搜索算法、搜索效果评价和问答系统等；此外，SIGIR 会议的主题中涉及用户研究的也比较多，如个性化推荐、情感分析、用户行为以及搜索体验和满意度等，总的来看，SIGIR 会议各年度论文主题推陈出新，如对于赞助商搜索广告的语义匹配研究[24]以及通过探索在线用户行为来提高个性化音乐检索效率[25]，信息检索的研究范畴正在逐渐扩大。

5　研究趋势分析

根据以上对热点词和主题词的分析以及对各年度会议论文的对比研究，笔者归纳出信息检索研究的如下发展趋势:

5.1　移动搜索成为主流研究内容

随着互联网和智能科技的不断发展，信息检索不再只有个人电脑终端(PC)搜索，用户越来越多地依赖移动设备来搜索他们所需的信息以及服务。从各会议的论文中发现:

(1) 移动搜索与周边的商家及服务密不可分。基于移动搜索和 PC 搜索的不同，Lv 等通过数据分析了用户在移动设备上的搜索日志[26]。

(2) 用户行为是目前整个学术界的研究热点，用户移动搜索行为也在信息检索领域中有所涉及。Lagun 等通过实验研究了用户在移动搜索中对移动设备视窗的注意情况[27]。

(3) 较多学者正在评估用户对移动搜索的满意度。Williams 等研究了搜索答案对于移动搜索用户的影响以及用户相应的满意度[28]。

5.2　信息检索模型正在优化和拓展

信息检索模型是信息检索的主要研究内容，其运用数学或其他语言与工具，对于信息检索的查询和文档及其匹配程度进行抽象描述，目前的信息检索模型包括布尔模型、向量空间模型、概率模型、语言模型以及基于本体的检索模型等[29]。各会议对于布尔模型的研究较少，而主要关注其他模型的优化和拓展应用。从这些相关论文中可以看出:

(1) 语言模型是当前最受关注的信息检索模型，无论是在社交网站上的应用，还是运用到多媒体检索或特定文档的检索任务。Tsagkias 等开发了推断用户浏览行为的语言动机模型[30]；Chen 等则将多峰语言模型用于演讲视频检索中[31]；Raviv 等研究了基于实体的语言模型在小说文档中的检索效果[32]。

(2) 概率模型则更多地用于提高检索的效率。Zhao 等构建了上下文相关的邻近检索模型以提升检索的效率[33]。

5.3　更加注重过滤与推荐

在信息检索中，过滤与推荐是满足用户信息需求的重要技术，有关过滤与推荐的算法和系统始终是学者们关注的问题。从各会议的论文中得出以下结论:

(1) 协同过滤是利用某兴趣相投、拥有共同经验的群体的喜好来推荐用户感兴趣的信息，其突出优点在于能够结合其他人的经验，过滤机器难以自动内容分析的信息，避免了内容分析的不完全或不精确[34]。Shih 等提出了一种提升协同过滤中评价较少项目的方法[35]。

(2) 学者们对社交媒体内容的推荐以及相关研究颇有兴趣。Hayashi 等在研究 Twitter 信息检索时开发了一种将主题抽取和信息流过滤融合的流算法并应用于 Twitter 信息流中[36]。

(3) 多媒体信息是当前信息检索领域中过滤与推荐研究关注的最热门内容。Lu 等研究了基于创新计算的在线视频推荐系统[37]；Mao 等以歌曲难度评级为依据开发了首个社交歌唱社区推荐系统[38]。

5.4　信息检索与人工智能关系密切

信息检索与计算机科学有着千丝万缕的联系，而当前人工智能是计算机科学的一个热门分支，因此信息检索与人工智能的关系相当密切。纵观各会议的论文主要有以下几个方面的发现:

(1) 机器学习应用于信息检索中的查询与鉴别。Lu 等运用非监督式学习的方法系统地鉴别音乐数据集中的异常部分[39]。

(2) 神经网络在跨语言检索中的应用。Zhou 等建立了一种弱共享深度神经网络结构来解决跨语言情感分类中源语言数据与目标语言数据的特征空间重叠问题[40]。

(3) 在机器学习中运用信息检索可以实现人机对话等多方面的功能。Yan 等运用信息检索、自然语言处理等技术建立了人机之间的自动对话系统[41]。

5.5　隐私问题在信息检索中广受关注

伴随着信息检索技术的发展，隐私的泄露和非法

交易等问题不断出现，因此隐私问题也是信息检索领域的众多学者热衷于研究的。从相关论文中可以得出:

(1) 信息检索广泛应用于社交媒体用户的隐私保护问题。Zerr 等提出一种自动检测隐私照片的技术并开发了基于隐私意识的照片分类检索系统[42]。

(2) 差分隐私(Differential Privacy)是基于数据失真的隐私保护技术，通过向查询或者分析结果中添加噪音使数据失真，从而达到隐私保护的目的[43]。Zhang 等介绍了一种利用差分隐私技术将查询日志匿名化的检索框架[44]。

(3) 信息检索还用于研究个性化网络搜索中的隐私问题。Ahmad 等提出了建立在客户端基于主题的隐私保护措施以解决搜索引擎保存用户历史搜索记录可能产生的隐私问题[45]。

5.6 医疗与健康成为焦点

近年来，医疗与健康问题已经成为各学科的研究热点，在信息检索领域也不例外。通读各会议相关论文发现信息检索对医疗与健康的研究主要集中在以下方面:

(1) 医疗信息检索的质量和效果是学者们所重点关注的。Schoenherr 等研究了挖掘健康信息的查询和搜索历史的潜在策略[46]。

(2) 通过信息检索可以实现对用户健康状况的监控。Sidana 等运用主题模型的思想提出了一种新的潜在疾病模型[47]。

6 结 语

本文通过对 5 个相关会议的论文进行分析与综述，发现了信息检索的研究现状及发展趋势，为相关学者展现了该领域的大致发展面貌，从而为开拓新的研究方向与主题提供了一定的借鉴意义。但由于部分会议论文的获取难度较大，仅能采集到题目、摘要和关键词等信息，无法从全文的视角进行更具深度的分析，因此下一步值得开展的研究是进行基于全文内容以及引文的分析，从而更好地了解信息检索的最新动态。

参考文献:

[1] 司莉, 庄晓喆, 贾欢. 近 10 年来国外多语言信息组织与检索研究进展与启示[J]. 中国图书馆学报, 2015, 41(4): 112-126. (Si Li, Zhuang Xiaozhe, Jia Huan. A Review of Multilingual Information Organization and Retrieval Research Abroad in the Last Ten Years[J]. Journal of the Library Science in China, 2015, 41(4): 112-126.)

[2] 吴丹, 邱瑾. 国外协同信息检索行为研究述评[J]. 中国图书馆学报, 2012, 38(6): 100-110. (Wu Dan, Qiu Jin. A Review on Foreign Studies of Collaborative Information Seeking Behavior [J]. Journal of the Library Science in China, 2012, 38(6): 100-110.)

[3] 杨海锋. 用户行为在信息检索中的研究现状及发展动态评述[J]. 图书情报知识, 2015(6): 79-88. (Yang Haifeng. Review on the Research Status and Development Trend of Users Behavior in Information Retrieval [J]. Document, Information & Knowledge, 2015(6): 79-88.)

[4] 窦永香, 苏山佳, 赵捧未. 信息检索研究的发展与动向——对 ACM SIGIR 信息检索年会的主题分析[J]. 情报理论与实践, 2010, 33(7): 124-128. (Dou Yongxiang, Su Shanjia, Zhao Pengwei. Progress and Development Trend in the Study of Information Retrieval[J]. Information Studies: Theory & Application, 2010, 33(7): 124-128.)

[5] 陈少涌, 李广建. 近十年来信息检索研究发展动向——基于 SIGIR 年会主题及论文集的统计分析[J]. 情报科学, 2015, 33(5): 150-156. (Chen Shaoyong, Li Guangjian. Research on Information Retrieval over the Last Decade: Analysis of SIGIR Annual Conferences' Research Topics and Proceedings[J]. Information Science, 2015, 33(5): 150-156.)

[6] Casey M A, Veltkamp R, Goto M, et al. Content-Based Music Information Retrieval: Current Directions and Future Challenges[J]. Proceedings of the IEEE, 2008, 96(4): 668-696.

[7] Kishida K. Technical Issues of Cross-language Information Retrieval: A Review [J]. Information Processing and Management, 2005, 41(3): 433-455.

[8] Enser P. The Evolution of Visual Information Retrieval[J]. Journal of Information Science, 2008, 34(4): 531-546.

[9] Smeaton A F, Keogh G, Gurrin C, et al. Analysis of Papers from Twenty-Five Years of SIGIR Conferences: What Have We been Doing for the Last Quarter of a Century?[J]. ACM SIGIR Forum, 2002, 36(2): 39-43.

[10] Hiemstra D, Hauff C, Jong F. SIGIR's 30th Anniversary: An Analysis of Trends in IR Research and the Topology of Its Community[J]. ACM SIGIR Forum, 2007, 41(2): 18-24.

[11] Christopher D, Surdeanu M, Bauer J, et al. The Stanford CoreNLP Natural Language Processing Toolkit [C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.2014.

[12] Park D C, El-Sharkawi M A, Marks R J, et al. Electric Load Forecasting Using an Artificial Neural Network[J]. IEEE Transactions on Power Engineering, 1991, 6(2): 442-449.

[13] Pineda F J. Generalization of Back-Propagation to Recurrent Neural Networks[J]. Physical Review Letters, 1987, 59(19): 2229-2232.

[14] Krizhevsky A, Sutskever, B, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[C]//Proceedings of the 2012 Advances in Neural Information Processing Systems.2012.

[15] Yann L, Bengio Y, Hinton G. Deep Learning[J]. Nature, 2015, 521: 436-444.

[16] Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection[J]. Journal of Machine Learning Research, 2003, 3(6): 1157-1182.

[17] Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques for Recommender Systems[J]. IEEE Computer Society, 2009, 42(8): 30-37.

[18] 董士海. 人机交互的进展及面临的挑战[J]. 计算机辅助设计与图形学学报, 2004, 16(1): 1-12. (Dong Shihai. Progress and Challenge of Human-Computer Interaction[J]. Journal of Computer-Aided Design & Computer Graphics, 2004, 16(1): 1-12.)

[19] Steuer J. Defining Virtual Reality: Dimensions Determining Telepresence[J]. Journal of Communication, 1992, 42(4): 73-93.

[20] Azuma R T.A Survey of Augmented Reality[J]. Teleoperators and Virtual Environments, 1997, 6(4): 355-385.

[21] Davis J, Costa V S, Peissig P, et al. Demand-Driven Clustering in Relational Domains for Predicting Adverse Drug Events[C]//Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK. 2012.

[22] Wang Y, Wipf D, Ling Q, et al. Multi-Task Learning for Subspace Segmentation [C]//Proceedings of the 32nd International Conference on Machine Learning, Lille, France. 2015.

[23] 任磊. 推荐系统关键技术研究[D]. 上海: 华东师范大学, 2012. (Ren Lei. Research on Some Key Issues of Recommender Systems [D]. Shanghai: East China Normal University, 2012.)

[24] Grbovic M, Djuric N, Radosavljevic V, et al. Scalable Semantic Matching of Queries to Ads in Sponsored Search Advertising[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.2016: 375-384.

[25] Cheng Z, Shen J, Hoi S. On Effective Personalized Music Retrieval by Exploring Online User Behaviors[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2016: 125-134.

[26] Lv Y, Lymberopoulos D, Wu Q. An Exploration of Ranking Heuristics in Mobile Local Search[C]//Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2012: 295-304.

[27] Lagun D, Hsieh C H, Webster D, et al. Towards Better Measurement of Attention and Satisfaction in Mobile Search[C]//Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2014: 113-122.

[28] Williams K, Kiseleva J, Crook A C, et al. Is This Your Final Answer? Evaluating the Effect of Answers on Good Abandonment in Mobile Search[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.2016: 889-892.

[29] 孙坦, 周静怡. 近几年来国外信息检索模型研究进展[J]. 图书馆建设, 2008(3): 82-85. (Sun Tan, Zhou Jingyi. The Review of Information Retrieval Models in Recent Years[J]. Library Development, 2008 (3): 82-85.)

[30] Tsagkias M, Blanco R. Language Intent Models for Inferring User Browsing Behavior [C]//Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval.2012: 335-344.

[31] Chen H, Cooper M, Joshi D, et al. Multi-modal Language Models for Lecture Video Retrieval[C]//Proceedings of the 22nd ACM International Conference on Multimedia.2014: 1081-1084.

[32] Raviv H, Kurland O, Carmel D. Document Retrieval Using Entity-Based Language Models[C]//Proceedings of the 39th International ACM SIGIR Conference on Research & Development in Information Retrieval.2016: 65-74.

[33] Zhao J, Huang J. An Enhanced Context-Sensitive Proximity Model for Probabilistic Information Retrieval [C]//Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. 2014: 1131-1134.

[34] Goldberg D, Nichols D, Oki B M, et al. Using Collaborative Filtering to Weave an Information Tapestry [J]. Communications of the ACM, 1992, 35(12): 61-70.

[35] Shih T Y, Hou T C, Jiang J D, et al. Dynamically Integrating

Item Exposure with Rating Prediction in Collaborative Filtering[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.2016: 813-816.

[36] Hayashi K, Maehara T, Toyoda M, et al. Real-Time Top-R Topic Detection on Twitter with Topic Hijack Filtering[C]// Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 417-426.

[37] Lu W, Chung F. Computational Creativity Based Video Recommendation[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2016: 793-796.

[38] Mao K, Fan J, Shou L, et al. Song Recommendation for Social Singing Community[C]//Proceedings of the 22nd ACM International Conference on Multimedia. 2014: 127-136.

[39] Lu Y C, Wu C W, Lu C W, et al. An Unsupervised Approach to Anomaly Detection in Music Datasets[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.2016: 749-752.

[40] Zhou G, Zeng Z, Huang J, et al. Transfer Learning for Cross-Lingual Sentiment Classification with Weakly Shared Deep Neural Networks [C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.2016: 245-254.

[41] Yan R, Song Y, Wu H. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System [C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.2016: 55-64.

[42] Zerr S, Siersdorfer S, Hare J, et al. Privacy-Aware Image Classification and Search[C]//Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval.2012: 35-44.

[43] Dwork C, McSherry F, Nissim K, et al. Calibrating Noise to Sensitivity in Private Data Analysis[C]//Proceedings of the 3rd Theory of Cryptography Conference. 2006: 265-284.

[44] Zhang S, Yang H, Singh L. Anonymizing Query Logs by Differential Privacy[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2016: 753-756.

[45] Ahmad W U, Wang H. Topic Model Based Privacy Protection in Personalized Web Search[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2016: 1025-1028.

[46] Schoenherr G P, White R W. Interactions Between Health Searchers and Search Engines[C]//Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2014: 143-152.

[47] Sidana S, Mishra S, Amer-Yahia S, et al. Health Monitoring on Social Media over Time[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2016: 849-855.

作者贡献声明:

杨超凡: 设计研究方案, 分析数据, 起草论文;
邓仲华: 提出研究思路, 论文最终版本修订;
彭鑫: 审核和清洗数据, 论文最终版本修订;
刘斌: 采集数据。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 杨超凡, 刘斌. conf data. zip. 近五年来信息检索的研究现状与发展趋势综述关联数据.

收稿日期: 2017-05-22
收修改稿日期: 2017-07-02

Review of Information Retrieval Research: Case Study of Conference Papers

Yang Chaofan Deng Zhonghua Peng Xin Liu Bin
(School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] This paper reviews conference papers on information retrieval, aiming to identify the research hotspots and development trends in this field. [**Coverage**] Papers published by ACL, ACMMM, ICML, KDD, and SIGIR from 2012 to 2016. [**Methods**] We first collected these papers' abstracts and keywords to process them with word segmentation package. Then, we analyzed these data with statistic tests. [**Results**] We found that mobile search was the most popular topic and the information retrieval models had been optimized. Filtering and recommending received more attention from the researchers. Information retrieval studies established close ties with artificial intelligence. User's privacy protection and health information retrieval were also popular. [**Limitations**] Only collected the abstracts and keywords. More research is needed to study the full texts and citations. [**Conclusions**] This paper presents the latest developments of information retrieval research.

Keywords: Information Retrieval Conference Papers Research Hotspots Development Trends

学术研究：意见领袖在社交媒体中的重要性

跟踪研究一个月内随机抽样的 30 万活跃用户的 Twitter 更新，结果显示社会媒体和社交网络的这个特殊角落并不像人们认为的那样平等和民主。事实上，《国际互联网营销和广告杂志》(*International Journal of Internet Marketing and Advertising*)"上发表的研究表明，信息流动分为两步(Two-Step Flow of Information)，少数用户产生了大部分的影响力，意见领导人跟随其他意见领袖，并在广泛的用户群体内有效地形成一个有影响力的社区，传播的信息随着日常用户共享、转发和重用，遵循某种权力分配。

科罗拉多大学广告、公共关系和媒体设计系 Harsha Gangadharbatla 和 Twitter 的软件工程师 Masoud Valafar 解释，有关信息传播，以及"口碑"如何影响民意和消费决策的理论非常多，媒体和社交媒体对个人和团体的影响也有很多研究。

其中一种理论被称为两步流理论(Two-Step Flow Theory)。该理论认为，意见领袖的意见很容易诱导大众就某一主题形成意见。同时，这些意见领袖本身也受到大众媒体的影响。这与一步流理论(One-Step Flow Theory)形成对比，在该理论下，人们直接受到大众媒体的影响。显然，无论是电视、广播、报纸还是网络，人们都会不断地暴露在大众媒体之下。但是，研究人员认为，这个意见实际上更有可能是在两步过程中形成。对社交媒体分享的意见尤其如此，同样也可能适用于传统媒体的环境——电视专家、报刊杂志、专栏作家等。

人们往往认为，随着 YouTube、Twitter、Instagram，以及其他的 Web 2.0 网站等新媒体的出现，信息和影响力的民主化开始显现。Gangadharbatla 和 Valafar 认为，事实并非如此，至少在 Twitter 的环境下并不是这样。社交媒体正在彻底改变用户和消费者获取信息、新闻和意见的方式，但与过去一样，仍然存在有很大影响力的人，也即意见领袖，他们可能是信息中心、新闻媒体，甚至是名人，是信息和意见的主要来源。

该文章认为："在社交媒体上与传统媒体中，信息传播的方式并没有什么不同，换句话说，即使是像 Twitter 和 Instagram 这样的民主环境，信息也主要通过意见领袖进行传播，更为重要的是，这些意见领袖与媒介上的其他舆论领袖联系起来，形成一个虚构的意见领袖，对社交媒体上信息以何种方式以及何种速度进行传播产生强大的影响。因此，从商业角度来说，瞄准虚拟社区的意见领袖进行推广将比在 Twitter 上广撒网接触人民群众更加有效。"

(编译自: https://www.sciencedaily.com/releases/2017/08/170810104849.htm)

(本刊讯)